NECAT-CLWE: A SIMPLE BUT EFFICIENT PARALLEL DATA GENERATION APPROACH FOR UNSUPERVISED AND SEMI-SUPERVISED NEURAL MACHINE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Many languages lack sufficient data to train qualitative translation systems, particularly those based on the cutting-edge neural machine translation architectures. Recently, it has been demonstrated that using an exact copy of the monolingual target data as the source data improves the quality of translation systems, allowing them to benefit from proper nouns and such similar words that do not require translation. However, using an exact copy of the target data contaminates the source data with terms in the target language that needs translation. As a result, we describe in this paper a similar but more effective parallel data generation approach for improving low-resource neural machine translation using named entity copying and approximate translations using cross-lingual word embedding (NECAT-CLWE). The work will be evaluated on the low resource English-Hausa neural machine translation.

1 INTRODUCTION

Since the dawn of time, language and communication have been important to human interactions. As a result, linguistic translation has played a crucial role in sociological and cultural achievements. To simplify and enable communication between different languages, a means is required. Because these means can be expensive or difficult to obtain, researchers have explored creating translations automatically. Neural Machine Translation (NMT) (Bahdanau et al., 2015; Vaswani et al., 2017), a deep learning-based approach to machine translation, is currently the state-of-the-art approach to automatic translation. The approach, though, requires a large amount of human-annotated parallel data to reach acceptable level of accuracy.

Due to the paucity of parallel data to train qualitative translation systems, many research works have explored using other available resources, e.g., in Transfer Learning (Zoph et al., 2016) and Language Modelling (Lample et al., 2018); or generating additional training data through back-translations (Sennrich et al., 2016; Abdulmumin et al., 2021) and forward translations (Zhang & Zong, 2016). While the latter approaches have been shown to be successful at improving translation accuracy, they are very expensive to implement and the quality of the additional data is also not assured. Currey et al. (2017); Burlot & Yvon (2018) proposed a simple and cost-effective approach for utilizing the monolingual target data by copying the target data into the source language . This has been shown to benefit the translation model especially where the words are proper nouns or words that do not require translation. This approach, though, will litter the source language with words that are not in its vocabulary and require to be translated.

This research work, therefore, aims to improve upon the baseline work of Currey et al. (2017). While the named-entities will be shared between the source and target sentences, the unknown words that require translation will be replaced with the nearest words in the vocabulary of the source language using a pre-trained cross-lingual word embedding (JP et al., 2021). Cross-lingual word embedding have been shown to generate a similar vector representation for similar words across two or more languages (Zhang et al., 2019).

2 RESEARCH OBJECTIVES AND SIGNIFICANCE

The aim of this study is to create an efficient parallel data generation technique based on named entity copying and approximate translation using cross-lingual word embedding. The following objectives are presented to achieve the stated aim.

1. To conduct an extensive review of current parallel data creation algorithms.
2. To develop a parallel data generation algorithm using named entity copying and approximate word translation with cross-lingual word embedding.
3. To evaluate the proposed approaches against currently data generation algorithms on English-Hausa low resource NMT.

Significance of the Study Due to the importance of having qualitative translation systems especially in languages that are categorized as low-resourced, the successful conduct of this research work will provide an efficient method for improving the performance of machine translation systems in low and even high resource languages. The work, being evaluated on English-Hausa NMT, will also help further develop research in the local Hausa language.

3 THE PROPOSED NECAT-CLWE PARALLEL DATA GENERATION APPROACH

Firstly, we get the monolingual data in each of the target and source languages, mix the two data and train a cross-lingual words embedding. Secondly, we copy all the named entity words into the generated source sentences at the same positions as in the target sentences. Then, for each word in the target sentences that require translation, we select and insert into the source sentences the nearest word in the source vocabulary using the pretrained cross-lingual words embedding. These steps are iterated until all the target sentences are paired with their pseudo-parallel, approximate translations.

Algorithm 1 Named-Entity Copying and Approximate Translation using Cross-Lingual Words Embedding

Input: Monolingual source and target data: $X = \{(x^{(u)})\}_{u=1}^{U}$ & $Y = \{(y^{(v)})\}_{v=1}^{V}$

procedure PARALLEL DATASET GENERATION
    Train a cross-lingual words embedding (CLWE) vector on a mixture of the source and target monolingual data;
    Build the vocabularies of each of the source and target monolingual datasets;
    repeat
        Tokenize the target sentence;
        Extract all named-entities;
        Copy the named-entities into the same position in the source sentences;
        For each word in the target sentence that require translation, select and insert into the same position in the source sentence the nearest word in the source vocabulary using the pretrained CLWE;
    until monolingual target sentences are exhausted;
end procedure

Output: pseudo-parallel data, $P^{'} = \{(x^{(v)}, y^{(v)})\}_{v=1}^{V}$

4 CONCLUSION AND FUTURE WORK

Due to the absence of sufficient parallel datasets for training machine translation models, several research works are proposing efficient approaches for utilizing other resources to improve the accuracy of these models. In this work, we aim to utilize the abundant monolingual target data to improve low resource neural machine translation using named-entity copying and word substitution with cross-lingual words embedding. We aim to implement the proposed approach and evaluate it against current data generation algorithms.

REFERENCES

Idris Abdulmumin, Bashir Shehu Galadanci, Abubakar Isa, Habeebah Adamu Kakudi, and Ismaila Idris Sinan. A Hybrid Approach for Improved Low Resource Neural Machine Translation using Monolingual Data. *Engineering Letters*, 29(4):1478–1493, nov 2021. URL http://www.engineeringletters.com/issues_v29/issue_4/EL_29_4_21.pdf.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.0473.

Franck Burlot and François Francois Yvon. Using Monolingual Data in Neural Machine Translation: a Systematic Study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 144–155, Brussels, Belgium, oct 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6315. URL https://www.aclweb.org/anthology/W18-6315.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. Copied Monolingual Data Improves Low-Resource Neural Machine Translation. In *Proceedings of the Second Conference on Machine Translation*, volume 1, pp. 148–156, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4715.

Sanjanasri JP, Vijay Krishna Menon, Soman KP, Rajendran S, and Agnieszka Wolk. Generation of Cross-Lingual Word Vectors for Low-Resourced Languages Using Deep Learning and Topological Metrics in a Data-Efficient Way. *Electronics*, 10(12), 2021. ISSN 2079-9292. doi: 10.3390/electronics10121372. URL https://www.mdpi.com/2079-9292/10/12/1372.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised Machine Translation Using Monolingual Corpora Only. In *International Conference on Learning Representations*, 2018. URL http://arxiv.org/abs/1711.00043.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 86–96, Berlin, Germany, 2016. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017.

Jiajun Zhang and Chengqing Zong. Exploiting Source-side Monolingual Data in Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1535–1545, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/d16-1160.

Mozhi Zhang, Keyulu Xu, Ken-ichi Kawarabayashi, Stefanie Jegelka, and Jordan Boyd-Graber. Are girls neko or shōjo? cross-lingual alignment of non-isomorphic embeddings with iterative normalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3180–3189, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1307. URL https://aclanthology.org/P19-1307.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1568–1575, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1163. URL https://www.aclweb.org/anthology/D16-1163.